International Journal of Theoretical & Applied Computational Intelligence
https://ijtaci.com

Research Article

Modeling Visual Attention for Enhanced Image and Video Processing Applications

Uzair Ishtiaq[1]*, Ajaz Khan Baig[2] and Zubair Ishtiaque[3]

[1]Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
[2]Department of Computer Sciences, IBADAT International University, Islamabad, Pakistan
[3] Department of Analytical, Biopharmaceutical and Medical Sciences, Atlantic Technological University, H91 T8NW Galway, Ireland
[*]Corresponding Author: Uzair Ishtiaq (uzair@um.edu.my)
https://orcid.org/0009-0005-2407-0163

Received: 22/7/2025; Accepted: 27/9/2025; Published: 3/10/2025
https://doi.org/10.65278/IJTACI.2025.11

Abstract: Human attention is naturally drawn to visually salient or distinct stimuli. However, identifying all potentially interesting targets in a scene can be computationally complex. Visual saliency plays a crucial role in this process by highlighting important regions either through bottom-up (stimulus-driven) or top-down (goal-driven) mechanisms. In a bottom-up approach, attention is guided by inherent visual properties of the stimulus, whereas in a top-down approach, it is influenced by the user's intent or task. Over the past decade, researchers have developed various methods and models to detect visual distinctiveness in images and video frames. In this paper, we have discussed visual attention modelling, which has demonstrated wide-ranging applications, including image and video quality assessment, video summarization (such as video skimming and key frame extraction), and more. The findings of this study suggest that these models are efficiency as they reduce the computational complexity. Moreover, they enhance the performance of higher-level tasks in computer vision and multimedia analysis.

Keywords: Psychology; Human attention; Visual attention; Scene analysis; Video processing

1. Introduction

According to William James, in his book “Principles of Psychology”, which he wrote in 1890, attention is well known concept to all of us. It refers to what our mind takes possessions of any one thing out of several others that we observe simultaneously, clearly and brightly. At the same moment, we are not paying

attention to others but concentrating on one thing of interest to us. Visual attention is a theoretical notion which corresponds to the aptitude depicted by a person to keep concentration over a particular target. This is non-linear in nature. It means that naturally a human being cannot concentrate on all portions of his conspicuity area of a frame. We can attend to one thing at a time, and our attention helps us make a decision about which direction to move our eyes. Human perception is dependent on three basic factors, which are attention, eye movement and memory. The spatial region around the center of gaze within which the target can be sensed in the first glimpse or fixation is known as the conspicuity area. Here, gaze is the center of a stable, intentional look, and fixation is the act of sustaining the gaze in a constant direction [1].

Previously, psychophysical measurement procedures were slow and complex; however, the researchers have presented procedures that assess the complete conspicuity area, encompassing full awareness of the target while the scene maintains its location. With such method the target can easily be resolved from its surroundings [2]. The distinctive property of an image or any frame of a video that how much distinct it is from its background, is known as visual saliency. According to human observer studies, it has been established that saliency is one of the most significant properties of interest, gaze allocation, and attention in a person when they are freely viewing and observing any static or dynamic image [3]. Image saliency is the practical concept of visual attention. Through image saliency we can find how much an image or an object is distinct from rest of its background. Saliency can be measured computationally.

Figure 1: Visual saliency

Figure 1 is showing the Visual Saliency of different images, the salient parts of the images are highlighted when these images are applied different saliency techniques.

Visual saliency is done in the following two ways:

- Bottom-up Visual Saliency
- Top-Down Visual Saliency

1.1 Bottom-Up Visual Saliency

Bottom-up visual saliency is a stimulus-driven signal, which means that the saliency is depends over the stimulus [4]. The factors effecting bottom-up image saliency include: color, luminance, orientation contrast, etc

- For example: Consider there is a green field. If we found any red object in that field, it will surely capture our attraction towards itself, which is in a bottom-up manner.

Figure 2 (a): Visual saliency

Figure 2 (b): Bottom-up visual saliency

Figure 2(a) shows some models of houses, where all the houses are blue except one, which is red in colour. Figure 2 (b) is showing the resultant visually salient image of the houses when a bottom-up visual saliency technique is applied on it. The red house can easily be figured out.

1.2 Top-Down Visual Saliency

Top-Down visual saliency is a user-driven signal, which refers that the saliency is depends on the user. The factors effecting top-down image saliency include: “how” and “where” the image tend to appear in the scene [1].

- For example: If we have a toy basket and we are looking into it for the cricket ball we will notice “how” and “where” all of our toys located, still continuously searching for the cricket ball, which is in a top-down manner.

Figure 3 (a): Actual image

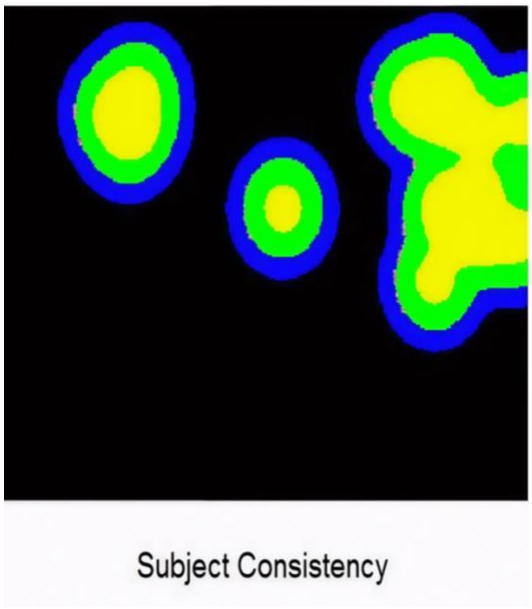

Figure 3 (b): Top-down image aliency

Figure 3 (a) presents the actual image in which the target is defined to be the paintings. In figure 3 (b) a top-down image saliency technique is applied on it and the resultant image is obtained in which the paintings are highlighted. Visual attention plays a critical role in autonomous vehicles by enabling them to focus on the most relevant parts of the driving environment, similar to how human drivers selectively attend to important cues on the road. Instead of processing every pixel in the scene equally, attention models help prioritize salient regions such as traffic signs, pedestrians, vehicles, lane markings, and obstacles. By combining bottom-up cues (e.g., motion, color contrast, unusual objects) with top-down cues (e.g., task-specific goals like detecting stop signs or monitoring crosswalks), these systems can make faster and more accurate decisions. This selective focus not only reduces computational load but also enhances safety and efficiency in complex, dynamic environments. Consequently, visual attention models are increasingly integrated into perception systems for autonomous driving, supporting tasks such as object detection, collision avoidance, and real-time navigation [5,36].

1.3 Models for Visual Saliency

To calculate visual saliency maps, numerous algorithms have been proposed over the past decade. With the use of these algorithms, a given input image can be converted into its consequent scalar-valued map and these maps are normally compared with human observer data for its validity. Researchers have

presented a number of computational saliency models, including, Attention-based on Information Maximization (AIM) [2], Saliency Using Natural Image Statistics (SUN) [3], Spectral Whitening (SW) [4], Edge Distance Saliency (EDS) [5], Saliency Toolbox 2.2 [6] [7], Extended Saliency (ESaliency) [8], Multiscale Contrast Conspicuity (MCC) [1], Rarity Based Saliency [9], Graph Based Visual Saliency (GBVS) [10], Phase Spectrum of Quaternion Fourier Transform (PQFT) [11], Frequency-Tuned Saliency (FTS) [12], Adaptive Whitening Saliency [13], etc.

There can be a lot of applications for these visual attention models, like, image and video quality assessment [14], video summarization [15], progressive image transmission [16], image segmentation [17], object recognition [18], etc.

2. Related Work

In this paper, the human visual conspicuity is first explained, along with the procedure for measuring conspicuity in humans. After that some computational saliency models are given.

2.1 Human Visual Conspicuity:

The area around the centre of gaze and attention from which the target can be identified in a single glance is known as visual conspicuity. If the target is implanted in the background of the image, then the conspicuity area would be small, whereas, it would be large if the target is clear and not embedded in its background. The conspicuity measurement procedure is given in the following steps [1]:

1- The observer fully observes the target to achieve full knowledge of its locality.
2- He marks a point that is at a large angular space from the point where the target as well as it is located in front of the parallel position to the target. This is the initial fixation.
3- Then the observer changes fixations repeatedly and thus gradually gets closer to the target until it is identified in its spatial region.
4- The angular distance between the initial fixation and target's center is recorded.
5- This process is repeated for three times.

Conspicuity estimates have the following two types, which can be given as:

- Detection Conspicuity
- Identification Conspicuity.

Let us go through these two estimates one by one.

2.1.1 Detection Conspicuity

It imitates the concept of bottom-up saliency in which the previous target fixation is not that much important for the detection of the target. It finds out how obviously the target area differs from its background area.

2.1.2 Identification Conspicuity

The top-down saliency concept is depicted in identification conspicuity, where the observer uses target features to identify the target that interests them. It is used to find out how obviously the image details show that whether it is the target or it can be extracted as the target.

2.2 Computational Saliency Models

Following are some of the computational saliency models that are used to find the distinctive property of an image or a video.

2.2.1 AIM: Attention-based on Information Maximization

In [2], authors presented a bottom-up attention model in such a way that they wanted to maximize the information from the sampled scene. The purpose of AIM is to convert the image feature plane into visual saliency map using Shannon's self-information measure. The core idea of AIM was that the saliency of an image feature is associated with its local surroundings. In [19] authors presented an image operator which is used to calculate the gaze in random natural scenes and landscapes when a human is freely viewing them, based on local information and statistics. In this technique, a scene is sampled in such a way that maximum information is obtained from the scene under consideration.

2.2.2 SUN: Saliency Using Natural Image Statistics

The SUN, Saliency Using Natural Image Statistics is used by local image features to calculate their bottom-up saliency. The core idea is to detect the important targets, i.e., the key features of the image. In SUN, both the saliency techniques, i.e. bottom-up and top-down saliencies are used. Bottom-up saliency is used in free viewing when the target is not specified, whereas top-down saliency is used when the target is specified [3].

2.2.3 RBS: Rarity-Based Saliency

1- Global Rarity: It takes the intensity of local image and calculates its mean and variance. [9]
2- Local Rarity: It takes the intensity of local image and calculates its contrast. [9]

2.2.4 EDS: Edge Distance Saliency

In EDS [5], the original image is first decomposed into edge-transformed images, which are then thresholded at different grey-scale levels, producing binary edge images. Now these binary edge images are then combined to produce the saliency map. Saliency and the pixel values in this map are not directly related but inversely related to each other, i.e. the pixel that is near the edge has lower value and the pixel that is away from the edge has comparatively higher value [31].

2.2.5 Saliency Toolbox 2.2

Saliency Toolbox 2.2 [6], [7] uses contrast, color and luminance maps and computes its saliency map. The core idea behind Saliency Toolbox 2.2 is that local contrast is important for the target.

2.2.6 GBVS: Graph-Based Visual Saliency

In Graph-Based Visual Saliency [10], dissimilarity feature maps are allocated distance-weights to calculate saliency. In this approach the image under consideration is converted into a specific representation that resembles a pyramid with three levels.

- At the first level intensity map is computed.
- Then the color map and then the orientation distribution map is calculated.
- Then a fully linked graph representation is created for these three feature maps.

After the creation of these above mentioned map representations, weights are allocated among different nodes that are not directly proportional rather they are inversely proportional to the similarity that was found in the feature values and their spatial distance among different nodes.

2.2.7 SW: Spectral Whitening

In SW (Spectral Whitening) [4], a map is constructed that focuses solely on salient features, ignoring non-informative background information that is of no use. This is an example of top-down saliency. This is similar to the common human eye phenomenon that we focus on the informative features of a scene and ignore the other non-informative features.

2.2.8 PQFT: Phase Spectrum of Quaternion Fourier Transform

Information of the interesting details is provided by the phase spectrum while considering any image. In this model of image saliency, phase spectrum of an image's quaternion Fourier transform is used to compute the saliency map [11]. This saliency map is generated by representing each pixel by a quaternion (a group of four) that comprises of color, intensity and motion features.

2.2.9 FTS: Frequency Tuned Saliency

In FTS (Frequency Tuned Saliency) [12], the target is fixated by the local image feature contrast. Contrast feature is used for color and luminance, which gives the concept of bottom-up saliency.

2.2.10 SDSR: Saliency Detection by Self-Resemblance

In SDSR [20], the target is focused by local feature contrast. The contrast matrix of the pixel under consideration is compared with the contrast matrices of the nearby pixels and their measured similarity is used to calculate the saliency map. SDSR is an example of bottom-up saliency map.

2.2.11 E-Saliency: Extended Saliency

In E-Saliency (Extended Saliency) [8], the image that is given as an input is fragmented and then the saliency is calculated as the global dissimilarity of those segments. According to this technique:

- Natural landscapes are made up of slam components having similar feature properties.
- If there are two or more segments that seem similar, they can all be either targets or non-targets.
- Total targets in a scene are a few.

2.2.12 MCC: Multiscale Contrast Conspicuity

For human fixation, intensity contrast is an important property. In MCC (Multi-scale Contrast Conspicuity) [1] saliency is computed based on intensity contrast of the target. Target contrast is the relation between the intensity of the supposed target and the surrounding area of the target including the target itself.

Therefore, for this first these two values will be calculated. The contrast of the surrounding area is calculated by progressively increasing the width of the target surround area. When the target and the total area, which includes the target and its surrounding area, are equal, the target contrast is equal to 1, as there is no surrounding area at all. And if the target and the total area which includes the target and surround area are not equal, that means that the value of contrast will be between greater than 0 and less than 1. Target contrast will increase if the surround area is minimized, conversely, it will decrease if surround area is maximized. MCC (Multi-scale Contrast Conspicuity) is an example of bottom-up saliency.

3. Discussion

In this section, different models are compared about which the researchers had used same data set for the measurement of their visual attention.

3.1 The Models

In SUN [3], the low-level features of an image are taken into account to detect the important feature of a given input image, low-level features of an image include, color, contrast, luminance, etc. [3] is used to find the bottom-up image saliency. SW [4] is used to construct a saliency map which only concentrates on the significant and informative elements of an image while ignoring the non-informative parts. In EDS [5], the actual image under consideration is first decomposed into edge transformed smaller images. These smaller images are given thresholds with respect to their grayscale levels which forms binary edge images. Then, these binary edge images are fused, and ultimately, a saliency map is formed. Pixel values of the saliency may be inversely related, which means that those pixels which are close to the edge will be having

lower values and the pixels which will be away from the edge will be having higher values. In MCC [1], the saliency is detected on the basis of target contrast at different scales. First the target intensity is calculated and then again intensity is calculated by gradually increasing the target surround area. Target intensity decreases when the surround area increases and vice versa.

3.2 Dataset

The data set used in this research is the TNO Human Factors Search_2 image dataset and is employed by several researchers reported in the literature [1,3,4,5]. These four models operate on single spatial domain as well as they are parameter-free. The images present in the dataset TNO Human Factors Search_2 is of very high resolution (6144 X 4096) pixels. These models sub-sampled this dataset images by a factor of 4 and the resulting images were of (1536 X 1024) pixels. The TNO Human Factors Search_2 dataset is an important resource used in research on visual perception, saliency, and target detection. It includes 44 high-quality color photographs of natural scenes, with each image containing one military vehicle as the target to be identified.

Table 1: The Models and their Detection Saliency

Models	Detection Saliency
SUN [3]	0.438
SW [4]	0.217
EDS [5]	0.443
MCC [1]	0.653

Table 1 is showing the Spearman's Rank Order Correlation of detection saliency exhibited by SUN [3], SW [4], EDS [5] and MCC [1].

Table 2: The Models and their Identification Saliency

Models	Identification Saliency
SUN [3]	0.702
SW [4]	0.522
EDS [5]	0.608
MCC [1]	0.843

Table 2 presents the Spearman's Rank Order Correlation of identification saliency exhibited by SUN [3], SW [4], EDS [5], and MCC [1].

Table 3: The Models and their Mean Search Time

Models	Mean Search Time
SUN [3]	0.735
SW [4]	0.398
EDS [5]	0.550
MCC [1]	0.735

Table 3 shows the mean search time of SUN [3], SW [4], EDS [5], and MCC [1]. Correlation of all of these saliency models (except EDS) over the target area are having maximum saliency values in accordance with these three aspects, which include, detection saliency, identification saliency as well as mean search time. When compared with human estimates, MCC Metric has the largest estimate.

In EDS, the mean output values are large which means that the target saliency is low and the maximal target values represent the internal structure of the target. If the maximal saliency is small then the target's internal structure is prominent and if it is large, then the target's internal structure is not articulate but the boundaries are clear. According to the results, the correlation between mean EDS values and human observer data is strong compared to the maximal EDS values. Therefore, the targets which are having less internal structure can be extracted as salient targets, whereas, those targets which are having more internal details are not that much salient, since the former is having clear boundaries.

The results are showing that the highest correlations are between identification saliency and maximal saliency of MCC Metric (which is 0.843). MCC is an efficient metric used to calculate bottom-up saliency, as it employs a simple conspicuity contrast approach. The results also depict that the highest correlation between the calculated maximum saliency and the mean search time is again MCC Metric (which is 0.735).

There are several aspects to imitate human visual system, like 1) distinctness of local image considered statistically (as found in SUN), when 2) contrast is taken into account (as found in SW) or when 3) edginess is under consideration (as found in EDS). Thus, bottom-up saliency can be expressed for these saliency models to find out what local feature differences are there, which may include the differences in color, texture, shape, size, luminance, etc. The detailed description is given in Algorithm 1.

Algorithm 1: Modeling Visual Attention for Enhanced Image & Video Processing

1. **Input:**
2. **Dataset:** D=TNO_Search_1, image size 6144x4096

Models $M=\{SUN,SW,EDS,MCC\}$

Subsample factor f=4

3. **Preprocessing:**

$$I_{resized} = \frac{I}{f}, size\ (I_{resized}) = 1536x1024$$

4. **Model Evaluation:**

For each model m ∈ M:

$$S_m = m(I_{resized})$$

5. **Compute metrics:**

$$DS_m = f_{DS}(S_m),\ IS_m = f_{IS}(S_m), MST_m = f_{MST}(S_m)$$

Where Detection Saliency (DS), Identification Saliency (IS), Mean Search Time (MST)

6. **Output:**

$$7.\ Return\ \{\ DS_m,\ IS_m, MST_m \mid m \in M$$

8. **End:**

4. Applications

Visual attention modeling has a number of applications, which include video summarization [15], image and video quality assessment [14], progressive image transmission [16], image segmentation [17], object recognition [18], etc.

4.1 Video Summarization

There is a massive data placed on the internet which includes images and videos that is needing to be retrieved in a well-organized manner. With the technological advancement, the quantity of producing new and new video data is also mounting quickly. There must be some ways and means to browse this video data efficiently. Earlier techniques used to select the key frames randomly or on the interval basis, that is, selecting the frames after a particular interval of time. This problem can be resolved in many ways; one solution is to provide summaries of the video, allowing users to browse through it quickly. Except browsing, the users can easily reach the portion of the video of their interest.

Video Summarisation is a technique in which a video is condensed into a smaller size, highlighting only the most important and visually salient parts of the video. In addition. it is a technique in computer vision that aims to generate a concise representation of a video by selecting its most informative and visually salient parts while discarding redundant or less important segments. The process relies on visual attention and saliency models to identify the key frames or shots that best capture the essential content, such as important objects, actions, or scene changes. Depending on the approach, video summarization can be static, where a set of keyframes is extracted to form a visual storyboard, or dynamic, where short clips are stitched together to create a condensed version of the video. This technique has wide applications in surveillance, sports highlights generation, video browsing, and content retrieval, as it significantly reduces the time and storage needed while preserving the meaningful information that a human viewer would naturally focus on. There are two ways to perform video summarization:

- Video Skimming
- Key Frame Extraction

Video summarization is an old concept and a lot of work has been done by researchers in this field. Ma and Zhang [21] performed video skimming with their motion attention based model. Then Ma et al. [22] extended this work and produced a collection of visual, linguist and auditory features in an open framework. Chernyak and Stark [23] presented a top-down approach as they took visual attention depending upon the observer. [24] estimated the visual attention with the use of Visual Attention Index (VAI) then they grouped the extracted frames by K-means algorithm into different clusters and at the end those frames with the highest values of Visual Attention Index (VAI) were selected. Peng and Xiaolin [25] used color histogram to cluster the frames and after that they chose such a frame from each cluster which was the most salient. [26] used a time controlled algorithm to group related frames together which gives the concept of clustering and then that frame was selected as the key frame from each cluster that was visually most prominent and distinct based on these features like, color, texture and motion.

In video skimming, the original video is cut-short into a much smaller video which is shorter in duration compared to the real video. Video skimming produces skims which are more meaningful and enjoyable in contrast to the key frames. Whereas in key frames extraction based technique, some visually salient or distinct frames are extracted from the real video. Key frames let the user experience the whole significant and important parts of the video in just one view [27] exclusive of even viewing a single small clip. Most of the video summarization techniques use low level index features [27].

In contrast, a better approach could have been to employ high-level semantic content, which includes objects, proceedings, and actions in the video. Some researchers [27] extracted those frames from the videos which were visually distinct from their background with the use of some visual attention modeling techniques. One way to achieve this is to extract the key frames non-linearly to identify static visual attention signs, while the others may be dynamic. Once these signs are identified, they can be merged. Temporal gradients are used for dynamic attention modeling and that image saliency discovery which is based on image signature is used for static modeling. In image signature based saliency discovery technique, the foreground of the image is approximated. The underlying hypothesis is that, the foreground of the image is more informative and prominent as compared to the background of the image.

There are two methods for attention detection, which are,

- Dynamic attention detection and
- Static attention detection.

According to dynamic attention detection, human's visual attention can normally be attracted by motion contrast [24]. In addition, dynamic attention detection, deals with videos or image sequences and incorporates temporal information to model human visual attention over time. In addition to static features like color and contrast, it considers motion, flicker, and changes in the scene that naturally draw human focus in dynamic environments. For example, a moving object in a video is more likely to attract attention than a static background, even if both share similar visual features. Dynamic attention detection is crucial for tasks such as video summarization, surveillance, human-computer interaction, and autonomous driving, where attention shifts continuously in response to motion and scene changes. By integrating both spatial and temporal cues, dynamic attention detection provides a more accurate and realistic model of human visual perception in real-world environments [34]. However, static attention detection focuses on identifying salient regions within a single image, without considering motion or temporal changes. It relies on low-level visual features, such as colour, intensity, texture, orientation, and spatial contrast, to determine which areas of the image are most likely to attract human attention. This type of attention detection is particularly useful in applications such as image compression, object recognition, and image retrieval, where the goal is to prioritize or preserve the most visually important regions of a static scene. Because static attention detection does not account for motion or time-based cues, it is mainly applied to still images or scenarios where temporal information is irrelevant [35].

It is a natural fact that the brain and Human Visual System coordinate with each other to identify the visually salient portions from the images and videos. Human beings are able of focusing on particular areas of the images and videos by keenly observing them which is a neurobiological process known as human attention. Some precise mechanism for human attention needs to be explored, as no one has yet explored it. Human attention is expressed through two mechanisms: the first is the bottom-up attention mechanism, and the other is known as the top-down attention mechanism. Top-down attention is derived from high-level features, such as objects, actions, and events, which attract the observer's gaze. Bottom-up attention, in contrast, is derived by low level features which include color, texture, motion contrast, etc.

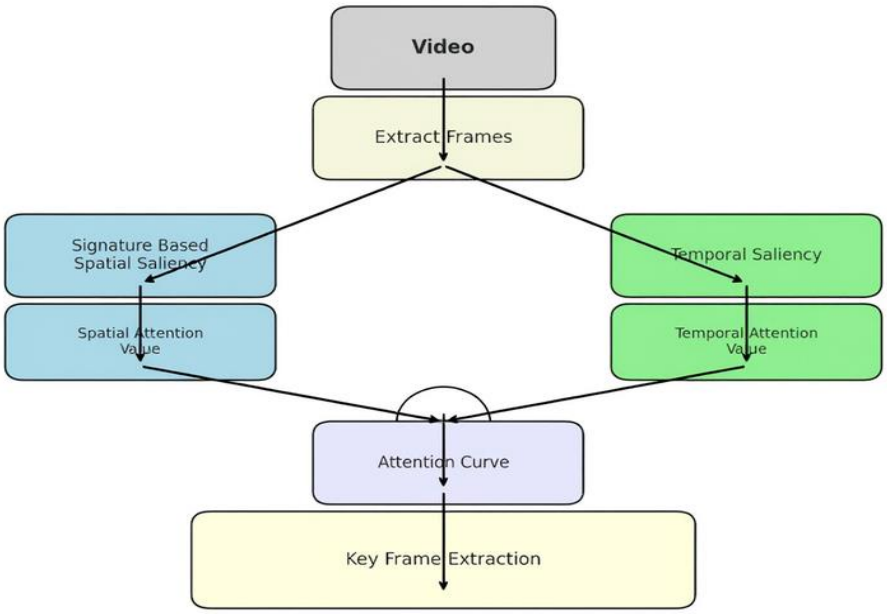

Figure 4: Key frame extraction

In the above architecture as illustrated in Figure 4, presented by Ejaz et al. [27], first of all frames are figured out from the video. Then first for dynamic frames, time based temporal saliency is used, resulting into temporal saliency attention values. Similarly, for static frames, image signature technique is used; resulting into signature based spatial saliency attention values. Then, these two found values are fused together, and an attention curve is obtained; thus, key frames are extracted that are visually salient.

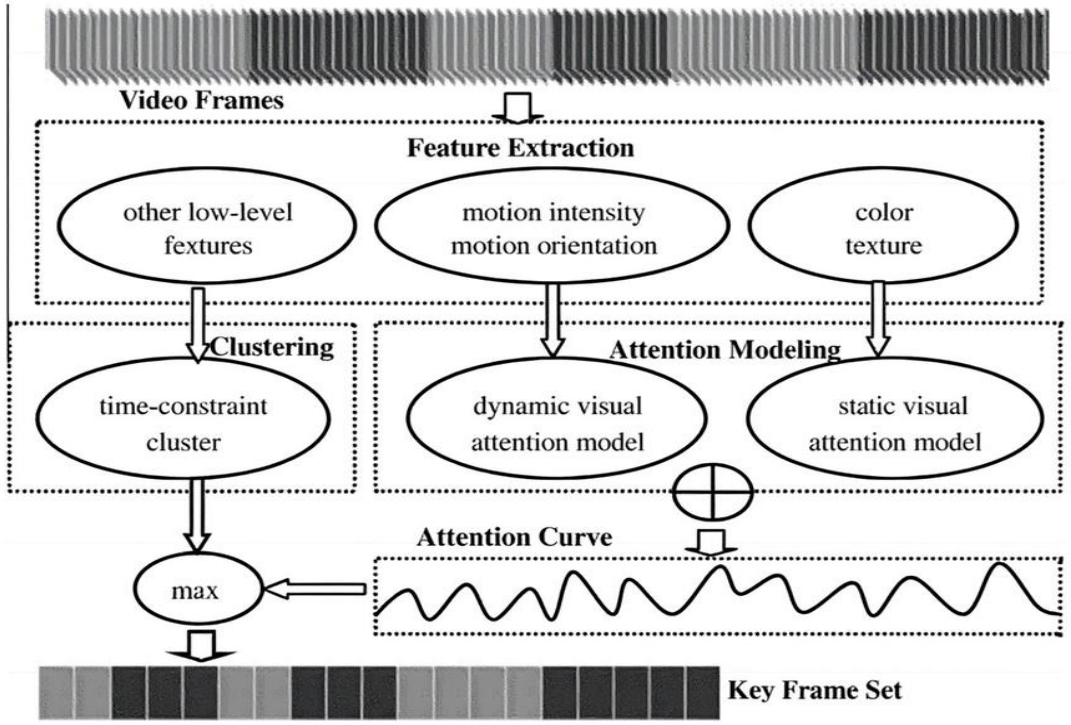

Figure 5: Lai and Yi et al. Architecture [26] for key frame detection

This architecture for key frames extraction as shown in Figure 5, was presented in [26] and is very much similar to Ejaz et al. [27]. In this architecture, first of all frames are extracted from the video based on their features, as mentioned here like color, texture, motion intensity, motion orientation and other low level features. They then utilised colour and texture features to create the static visual attention model, whereas motion intensity and motion orientation are employed for dynamic visual attention modelling. These two attention models (dynamic and static) are then merged through some fusion technique to achieve an attention curve. On other hand, frames with similar contents are clustered to create time-constraint clusters [26]. Now the attention curve and the time constraint cluster are used in a function named "max" which is used to get the final set of visually distinct key frames. Figure 6 is showing the extracted key frames by different techniques for a video.

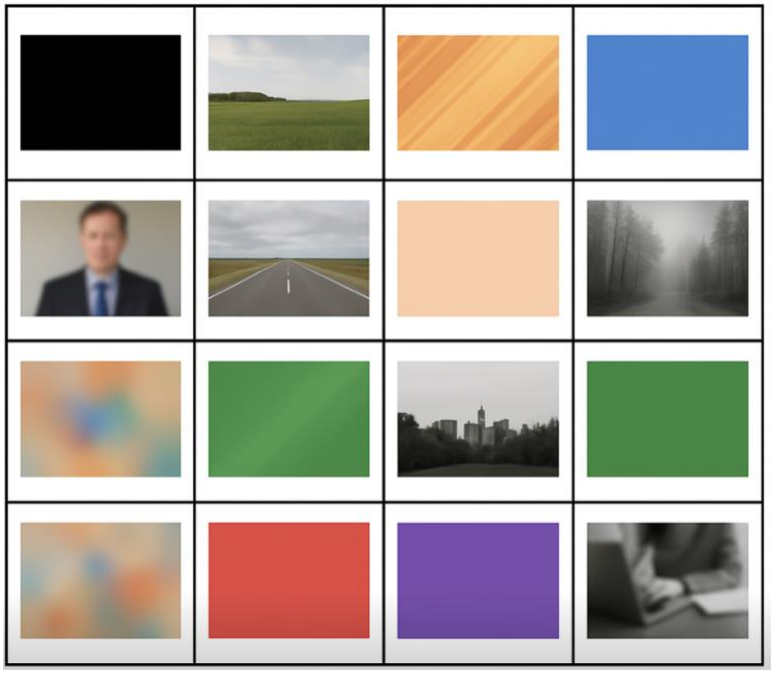

Figure: 6 Extracted key frames by different techniques

4.2 Image and Video Quality Assessment

In those applications in which images and videos are processed, their quality evaluation and assessment are one of the most significant concerns. There are a lot of indices that are used to cater human visual sensitivity which include the followings [14,27]:

- Structural Similarity (SSIM),
- Visual Information Fidelity (VIF)

For Structural Similarity (SSIM), when we are observing any scene, in a video or an image, we do observe luminance in it. Luminance is essentially the product of illumination that falls on objects and their reflectance. So, the structural information of the object is absolutely does not depend on the intensity of luminance and that of contrast. To extract that information, structure is separated from the influence of the intensity of luminance and then the SSIM Index is calculated [14]. Structural Similarity (SSIM) Index is

used for quality assessment. Applying Structural Similarity (SSIM) Index locally rather than globally is a better approach. It has a number of benefits. First, the statistic features, such as the luminance and the contrast, of natural images are not spatially stationary and they keep on changing. Second, image distortions may also be spatially different. Third, Human Visual System (HVS) is capable of foveation in which the image resolution varies all over the image as there are normally more than one fixation points, humans perceive only such particular area in an image which is having the highest resolution and is catered by the retina (also known as the fovea). Additionally, we obtain more information regarding image degradation through local quality measurements, which can help assess different applications. The Structural Similarity Index (SSIM) is a widely used perceptual metric for image quality assessment that evaluates the similarity between a reference image and a distorted version. Unlike traditional error-based measures, such as Mean Squared Error (MSE) or Peak Signal-to-Noise Ratio (PSNR), SSIM is designed to model how human vision perceives images by focusing on structural information rather than pixel-wise differences. It considers three key components: luminance, contrast, and structure, which together reflect how humans interpret image content. By emphasizing structural fidelity, SSIM provides a more meaningful measure of visual quality, making it especially useful in applications such as image compression, transmission, and restoration, where perceptual quality is more important than exact pixel-level accuracy. [32]

For Visual Information Fidelity (VIF), the correctness of the image is measured. Most of the researchers have concentrated to measure the signal fidelity for the assessment of visual quality. A number of models can be used to determine image and video quality assessment, like [2,11], etc. Visual Information Fidelity (VIF) is a more advanced image quality metric that measures the amount of visual information from a reference image that is preserved in a distorted image. Based on an information-theoretic framework, VIF relies on natural scene statistics and models the human visual system to estimate how much mutual information can be extracted from the distorted image relative to the original. In other words, it evaluates not just how similar two images look, but how much meaningful a human observer can still perceive content or information. This makes VIF particularly valuable in scenarios such as video compression, image transmission, and visual quality enhancement, where retaining essential perceptual information is critical even if some fine details are lost [33].

A lot of research has been done to simulate the Human Visual System (HVS) by the researchers [28]. The underlying principle of visual quality assessment is that an object present in the salient region of the image may be more disturbing than it would be in other non-informative parts. This can be found by recording the eye movements.

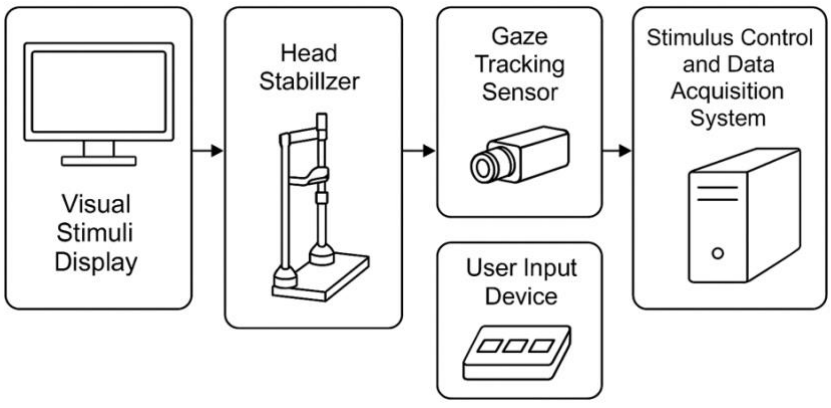

Figure 7: Eye tracking apparatus

5. Conclusion

Visual saliency is a non-linear concept and an important property of an image that determines which regions stand out and naturally attract human attention. It plays a crucial role in visual perception by highlighting the most informative or noticeable parts of an image. Broadly, saliency can be categorized into two types: bottom-up and top-down. Bottom-up saliency is stimulus-driven, meaning it is guided by low-level image features such as color, intensity, contrast, orientation, and motion. In contrast, top-down saliency is user-driven and influenced by prior knowledge, specific tasks, or contextual information; for example, when a person searches for a particular object in an image or video. To model these processes computationally, visual attention models are employed. These models attempt to simulate human visual perception by identifying the most distinct or relevant areas in an image or video. They have been widely applied in various real-world scenarios, including video summarization (selecting the most important frames or segments), feature-based visual attention (focusing on key regions for object detection or recognition), and image and video quality assessment (evaluating perceptual quality based on salient regions). Such models not only improve efficiency by reducing computational complexity but also enhance the performance of higher-level tasks in computer vision and multimedia analysis.

Funding: No specific funding received for this research.

Data Availability: The data that support the findings of this study is reported in section 3.4 of this research.

Conflicts of Interest: No conflict of interest is stated by the author.

Authors contributions. Conceptualization: UI, ZI; methodology: UI, AKB; validation: AKB, ZI; writing—original draft preparation, UI, AKB, ZI; writing—review and editing: UI, AKB, ZI; visualization: UI, ZI; supervision: UI, ZI; project administration: UI. The author had approved the final version.

References

[1] Chen, Qili, Junfang Fan, and Wenbai Chen. (2021) "An improved image enhancement framework based on multiple attention mechanism" *Displays* 70, 102091.

[2] Denison, R. N. (2024). "Visual temporal attention from perception to computation". *Nature Reviews Psychology*, 3(4), 261-274.

[3] Liu, Y., Dong, X., Zhang, D. et al., (2024). "Deep unsupervised part-whole relational visual saliency". *Neurocomputing*, 563, 126916.

[4] Cheng, S., Lu, Q., Shen, Z., et al., (2024). "3D Pop-Ups: Omnidirectional image visual saliency prediction based on crowdsourced eye-tracking data in VR". *Displays*, 83, 102746.

[5] Rosin, P. L. (2009). "A simple method for detecting salient regions". *Pattern Recognition*, 42(11), 2363-2371.

[6] Sharif, U., Mehmood, Z., Mahmood, T., et al., (2019). "Scene analysis and search using local features and support vector machine for effective content-based image retrieval". *Artificial Intelligence Review*, 52, 901-925.

[7] Walther, D., and Koch, C. (2006). "Modeling attention to salient proto-objects". *Neural Networks*, 19(9), 1395-1407.

[8] Avraham, T., and Lindenbaum, M. (2010). "Esaliency (extended saliency): Meaningful attention using stochastic image modeling". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4), 693-708.

[9] Lu, X., Jian, M., Wang, X., Yu, H., et al., (2022). "Visual saliency detection via combining center prior and U-Net". *Multimedia Systems*, *28*(5), 1689-1698.

[10] Harel, J., Koch, C., and Perona, P., "Graph-based visual saliency", *Advances in Neural Information Processing Systems 19*, The MIT Press, 545-552, (2007).

[11] Chenlei Guo, and Liming Zhang (2010). "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression". *IEEE Transactions on Image Processing*, 19(1), 185-198.

[12] Achanta, R., Hemami, S., Estrada, F., et al., "Frequency-tuned salient region detection". In 2009 IEEE Conference on Computer Vision and Pattern Recognition: IEEE, (2009).

[13] Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., et al., "Saliency based on decorrelation and distinctiveness of local responses", *Lecture Notes in Computer Science*, Berlin, Heidelberg: Springer Berlin Heidelberg, 261-268, (2009).

[14] Ma, Q., Zhang, L., and Wang, B. (2010). "New strategy for image and video quality assessment". *Journal of Electronic Imaging*, 19(1), 011019-011019.

[15] Guironnet, M., Pellerin, D., Guyader, N., et al., (2007). "Video summarization based on camera motion and a subjective evaluation method". *Eurasip Journal on Image and Video Processing*, 2007, 1-12.

[16] Rodríguez-Sánchez, R., Fdez-Valdivia, J., Toet, A., et al., (2004). "The relationship between information prioritization and visual distinctness in two progressive image transmission schemes". *Pattern Recognition*, 37(2), 281-297.

[17] Ko, B. C., and Nam, J. (2006). "Object-of-interest image segmentation based on human attention and semantic region clustering". *Journal of the Optical Society of America A*, 23(10), 2462.

[18] Walther, D., Rutishauser, U., Koch, C., et al., (2005). "Selective visual attention enables learning and recognition of multiple objects in cluttered scenes". *Computer Vision and Image Understanding*, 100(1-2), 41-63.

[19] Bruce, N. D. (2005). "Features that draw visual attention: An information theoretic perspective". *Neurocomputing*, 65-66, 125-133.

[20] Sun, Y., Min, X., Duan, H., et al., (2023, May). "The influence of text-guidance on visual attention". *In 2023 IEEE International Symposium on Circuits and Systems (ISCAS)* (pp. 1-5). IEEE.

[21] Sharif, U., Mehmood, Z., Mahmood, T., et al., (2019). "Scene analysis and search using local features and support vector machine for effective content-based image retrieval". *Artificial Intelligence Review*, 52(2), 901-925.

[22] Ma, Y. F., Hua, X. S., Lu, L., et al., (2005). "A generic framework of user attention model and its application in video summarization". *IEEE Transactions on Multimedia*, 7(5), 907-919.

[23] Chernyak, D., and Stark, L. (2001). "Top-down guided eye movements". *IEEE Transactions on Systems, Man and Cybernetics, Part B (cybernetics)*, 31(4), 514-522.

[24] Mehmood, Z., Gul, N., Altaf, M., et al., (2018). "Scene search based on the adapted triangular regions and soft clustering to improve the effectiveness of the visual-bag-of-words model". *Eurasip Journal on Image and Video Processing*, 2018(1).

[25] Peng, J., and Xiaolin, Q. (2009). "Keyframe-based video summary using visual attention clues". *IEEE Multimedia*, 11(11), 64-73.

[26] Lai, J., and Yi, Y. (2012). "Key frame extraction based on visual attention model". *Journal of Visual Communication and Image Representation*, 23(1), 114-125.

[27] Ejaz, N., Mehmood, I., and Wook Baik, S. (2013). "Efficient visual attention based framework for extracting key frames from videos". *Signal Processing: Image Communication*, 28(1), 34-44.

[28] Ninassi, A., Le Meur, O., Le Callet, P., et al., "Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric". In 2007 IEEE International Conference on Image Processing: IEEE II - 169-II - 172, (2007).

[29] Itti, L. (2007). "Visual salience". *Scholarpedia*, 2(9), 3327.

[30] Cheng, M., Zhang, G., Mitra, N. J., et al., "Global contrast based salient region detection". In Cvpr 2011: IEEE 409-416, (2011).

[31] Hassanin, M., Anwar, S., Radwan, I., et al., (2024). "Visual attention methods in deep learning: An in-depth survey". *Information Fusion*, 108, 102417.

[32] Mei, W., He, K., Xu, D., et al., (2025). "A lightweight medical image fusion network by structural similarity pseudo labels iteration". *Biomedical Signal Processing and Control*, 110, 108043.

[33] Moore, M. J., Robinson, A. K., and Mattingley, J. B. (2024). "Expectation modifies the representational fidelity of complex visual objects". *Imaging Neuroscience*, 2.

[34] Eskandari Nasab, M., Raeisi, Z., Lashaki, R. A., et al., (2024). "A GRU-CNN model for auditory attention detection using microstate and recurrence quantification analysis". *Scientific Reports*, 14(1), 8861.

[35] Zhou, W., and Li, X. (2024). "Pea-yolo: A lightweight network for static gesture recognition combining multiscale and attention mechanisms". *Signal, Image and Video Processing*, 18(1), 597-605.

[36] Liao, H., Li, Y., Li, Z., et al., (2024). "A cognitive-based trajectory prediction approach for autonomous driving". *IEEE Transactions on Intelligent Vehicles*, 9(4), 4632-4643.